ijecs open access www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 7 July, 2014 Page No. 6871-6875

A Modified *K*-Medoid Method to Cluster Uncertain Data Based on Probability Distribution Similarity

Aliya Edathadathil[1], Syed Farook [2], Balachandran KP[3]

[1]MTech in computer science and engineering,
MES Engineering College, Kuttippuram, Kerala
aliyaedathadathil@gmail.com

[2]Asst.Prof in Computer science and engineering,
MES Engineering College, Kuttippuram, Kerala
Ssyed.dimple@gmail.com

[3]Department of MCA,
MES Engineering College, Kuttippuram, Kerala
kpbala@gmail.com

Abstract: ***Clustering on uncertain data, one of the essential tasks in data mining. The traditional algorithms like K-Means clustering, UK-Means clustering, density based clustering etc, to cluster uncertain data are limited to using geometric distance based similarity measures, and cannot capture the difference between uncertain data with their distributions. Such methods cannot handle uncertain objects that are geometrically indistinguishable, such as products with the same mean but very different variances in customer ratings [6]. In the case of K-medoid clustering of uncertain data on the basis of their KL divergence similarity, they cluster the data based on their probability distribution similarity. Several methods have been proposed for the clustering of uncertain data. Some of these methods are reviewed. Compared to the traditional clustering methods, K-Medoid clustering algorithm based on KL divergence similarity is more efficient. This paper proposes a new method for making the algorithm more effective with the consideration of initial selection of medoids.***

Keywords: Uncertain data clustering, Probability distribution, *KL* divergence, Initial medoid.

1. Introduction

Clustering is one of the most important research areas in the field of data mining. In simple words, clustering is a division of data into different groups. Data are grouped into clusters in such a way that data of the same group are similar and those in other groups are dissimilar. Clustering is a method of unsupervised learning. Uncertainty in data arises naturally due to random errors in physical measurements, data staling, as well as defects in the data collection models. The main characteristics of uncertain data are, they change continuously, we cannot predict their behavior, the accurate position of uncertain objects is not known and they are geometrically indistinguishable. Because of these reason it is very difficult to cluster the uncertain data by using the traditional clustering methods .Clustering of uncertain data has recently attracted interests from researchers. This is driven by the need of applying clustering techniques to data that are uncertain in nature, and a lack of clustering algorithms that can cope with the uncertainty.

For example, in a shop the users are asked to evaluate a camera on the basis of various aspects such as quality, battery performance, image quality etc. Each camera may be scored by many users. Thus, the user satisfaction to a camera can be modeled as an uncertain object. There are often a good number of cameras under a user study. A frequent analysis task is to cluster the cameras according to user satisfaction data.

This paper is organized as follows: Section 2 presents a literature survey of the related works done in the field uncertain data clustering. Section 3 presents modified *K*-medoid algorithm to cluster uncertain data based on *KL* divergence. Section 4 shows the experimental results. Finally section 5 concludes this work.

2. Related Works

Various algorithms have been proposed for the clustering of uncertain data. Researchers are always trying to improve the performance and efficiency of the clustered data

Dr. T. Velmurugan.[1] proposed a *K*-means algorithm to cluster the data. Here the given set of data is grouped into *K* number of disjoint clusters, where the value of *K* is to be fixed in advance. The algorithm consists of two separate phases: the first phase is to define *K* initial centroids, one for each cluster. The next phase is to take each point belonging to the given data set and associate it to the nearest centroid. Generally Euclidean distance is used as the measure to determine the distance between data and the centroids. Then the centroids are

recalculated and clustering is done with these new centroids. The process is repeated till the clusters are not changed . *K*-means method is not that much efficient to cluster the uncertain data, that is the main disadvantage.

Samir N. Ajani[2] proposes an improved *K*-means algorithm to cluster the uncertain data effectively called *UK* means (Uncertain *K* means) clustering algorithm. *UK*-means basically follows the well-known *K*-means algorithm except that it uses Expected distance (ED) calculation instead of using Euclidian distance. Initially, k arbitrary point's $c_1 . . . c_k$ are chosen as the cluster representatives. Then, *UK*-means repeats the following steps until the result converges. First, for each data d_i, Expected Distance (d_i, c_j) is computed for all centroids and data. Data d_i is then assigned to cluster c_j that minimizes the Expected Distance. Here the computation of Expected distance involves numerically integrating functions, it is difficult to calculate.

The Expected distance calculation is one of the main problem in *UK*-means clustering. The efficiency of *UK*-means clustering is improved if the ED calculation is reduced. Ben Kao [3] proposes a new method to reduce the ED calculation in *UK*-means method. Here cluster the uncertain data by using *UK* means method with voronoi diagrams. Voronoi diagram divides the space into k cells V(cj) with the following property 1:

$$D(x, c_p) < D(x, c_q) \quad (1)$$

In each iteration, first construct the voronoi diagram from the *k* cluster representative points. For each data d_i, check if MBR lies completely inside any voronoi cell $V(c_j)$. If so, the data is assigned to cluster. In this case, no ED is computed. Here the effectiveness is not guaranteed, as it depends on the distribution of data.

Martin Ester.[4] proposes a density based clustering(DB Clustering) for clustering the data. The main difference of DB clustering is that here we do not need to specify total number of clusters in advance. Here to find a cluster, DB clustering starts with an arbitrary point p and retrieves all points density reachable from p wrt. NEps (Neighbourhood points with maximum radius) and MinPts (Minimum number of points in an NEps neighbourhood). If p is a core point, this procedure yields a cluster wrt. NEps and MinPts. If p is a border point, no points are density reachable from p and DB visits the next point of the database [5]. Need to specify NEps and MinPts, which can be difficult in practice.

2.1 *K*-medoid method for uncertain data clustering based on *KL* divergence similarity

Bin Jiang and Jian Pei.[6] proposed a new method for clustering uncertain data based on their probability distribution similarity. The previous methods extend traditional partitioning clustering methods like *K*-means, *UK* means and density-based clustering methods to uncertain data, thus rely on geometric distances between data. Probability distributions, which are essential characteristics of uncertain objects. Here systematically model uncertain objects in both continuous and discrete domains, where an uncertain object is modelled as a continuous and discrete random variable, respectively. Then use the well-known Kullback-Leibler (KL) divergence to measure similarity between uncertain objects in both the continuous and discrete cases, and integrate it into K-medoid method to cluster uncertain data. Compared to the traditional clustering methods, *K*-Medoid clustering algorithm based on *KL* divergence similarity is more efficient.

2.1.1 Uncertain Objects and Probability Distributions

Consider an uncertain object as a random variable following a probability distribution. We consider both the discrete and continuous cases. If the data is discrete with a finite or countable infinite number of values, the object is a discrete random variable and its probability distribution is described by a probability mass function (pmf). Otherwise, if the domain is continuous with a continuous range of values, the object is a continuous random variable and its probability distribution is described by a probability density function (pdf). For example, the domain of the ratings of cameras is a discrete set and the domain of temperature is continuous real numbers.

For discrete domains, the probability mass function of an uncertain data can be directly estimated by normalizing the number of observations against the size of the sample. The pmf of data P is expressed in eq.2

$$P(X) = \sum P_x(x)\, d_x \quad (2)$$

For continuous domains, the probability density function of an uncertain data can be calculated by using the following eq.3

$$P(X) = \int P_x(X)\, d_x \quad (3)$$

2.1.2 *KL* Divergence

After finding the probability distribution we have to find the probability distribution similarity between the data. Kullback-Leibler divergence (*KL* divergence) is one of the main method to calculate the probability distribution similarity between the data [7]. We show that distribution differences cannot be captured by the previous methods based on geometric distances. We use *KL* divergence to measure the similarity between distributions, and demonstrate the effectiveness of KL divergence using *K*-medoid clustering method.

In the discrete case, let f and g are two probability mass functions in a discrete domain with a finite or countably infinite number of values. The Kullback-Leibler diverge between f and g is defined in eq.3

$$D(f\|g) = \sum f(x) \log \frac{f(x)}{g(x)} \quad (3)$$

In the continuous case, let f and g be two probability density functions in a continuous domain with a continuous range of values. The Kullback-Leibler divergence between f and g is defined in eq.4

$$D(f\|g) = \int f(x) \log \frac{f(x)}{g(x)}\, dx \quad (4)$$

2.1.3 Applying *KL* divergence into *K*-medoid algorithm

K-medoid is a classical partitioning method to cluster the data[8]. A partitioning clustering method organizes a set of uncertain data into *K* number of clusters. Using *KL* divergence as similarity, Partitioning clustering method tries to partition data into *K* clusters and chooses the *K* representatives, one for each cluster to minimize the total *KL* divergence. *K*-medoid method uses an actual data in a cluster as its representative. Here use *K*-medoid method to demonstrate the performance of clustering using *KL* divergence similarity. The K-medoid method consists of two phases, the building phase and the swapping phase.

Building phase. In the building phase, the *K*-medoid method obtains an initial clustering by selecting initial medoids randomly.

Algorithm of building phase

Step 1. Randomly choose k number of data as the initial cluster.
Step 2. Calculate the *KL* divergence similarity between each medoids and the data.
Step 3. Assign the data to the medoid which has the smallest *KL* divergence with the medoid.
Step 4. Do swapping phase

Swapping Phase. In the swapping phase the uncertain *k*-medoid method iteratively improves the clustering by swapping a no representative data with the representative to which it is assigned.

Algorithm of swapping phase

Step1. Swapping the medoids with the non representative data.
Step2. Repeat step 2, 3 and 4 of building phase until the clusters are not changed.

Clustering uncertain data based on their probability distribution similarity is very efficient clustering method compare to other methods. But in the building phase the algorithm select initial medoids randomly that affect the quality of the resulting clusters and sometimes it generates unstable clusters which are meaningless. Also here the initial partition is based on the initial medoids and the initial partition affect the result and total number of iterations. If the initial medoids are selected in an efficient way then it does not produce any empty clusters and also we can reduce the total number of iterations.

3. PROPOSED METHOD

In the building phase of original *K*-medoid method the initial medoids are selected randomly that affect in the total number of iterations and sometime it generates meaningless empty clusters. To avoid these problems the proposed method modifies the building phase of the original *K*-medoid method by selecting the initial medoids effectively.

Algorithm of modified *K*-medoid clustering based on *KL* divergence method

Input:
$D = \{d_1, d_2..., d_n\}$ set of *n* data items.
K, Number of desired clusters

Output:
A set of *K* clusters.

Steps:
Phase 1: Determine the initial medoids of the clusters by using Algorithm 1.
Phase 2: Assign each data point to the appropriate clusters by using Algorithm 2.

Algorithm 1

Input:
$D = \{d_1, d_2,.....,d_n\}$ set of n data items
K = number of desired cluster

Output:
A set of *K* initial medoids

Steps:
1. Set p = 1
2. Compute the probability distribution similarity between each data and all other data in the set D
3. Find the most similar pair of data from the set D and form a data set Am which contains these two data , Delete these two data from the set D
4. Find the data in D that is similar to the data set A_m, Add it to Am and delete it from D
5. Repeat step 4 until the number of data in A_m reaches 0.75*(n/k)
6. If p<k, then p = p+1, find another pair of data from D between which the highest similarity, form another data set Am and delete them from D, Go to step 4
7. For each data set Am find the arithmetic mean of the vectors of data in Am, these means will be the initial medoids.

Algorithm 1 describes the method for finding initial medoid of the clusters effectively. Initially, compute the probability distribution similarity between each data and all other data in the set of data. Then find out the most similar pair of data and form a set A_1 consisting of these two data, and delete them from the data set D. Then determine the data which is similar to the set A_1, add it to A_1 and delete it from D. Repeat this procedure until the number of elements in the set A1 reaches a threshold. At that point go back to the second step and form another data set A_2. Repeat this till '*K*' such sets of data are obtained. Finally the initial medoids are obtained by averaging all the data in each data set. The *KL* divergence method is used for determining the probability distribution similarity between each data.

These initial medoids are given as input to the second phase , for assigning data to appropriate clusters. The steps involved in this phase are outlined as Algorithm 2.

Algorithm 2

Input:
$D = \{d_1, d_2,…...,d_n\}$ set of n data items
A set of k initial medoids

Output:
A set of k clusters

Steps:
1. Associate each data point to the most similar medoid. ("Similar" here is defined using *KL* divergence)
2. For each medoid m for each non medoid data o swap m and o and compute the total cost of the swapping

3. Select the swapping with the lowest cost
4. Repeat steps 2 to 5 until the clusters are not changed

4. EXPERIMENTAL RESULTS

For the experimental study, here taken around 104 cameras and each camera is rated by 100 users. The user satisfaction to a camera can be modeled as an uncertain object. The clustered result on the basis of users rating similarity can be used for the creation of recommendation system for e-commerce purpose. Implementations were done in Java. From the experimental results it can be seen that, the uncertain *K*-Medoid method generates unstable and empty clusters. And for each run the total number of iterations corresponds to each cluster is very high. But in the case of modified *K*-Medoid method it does not generate any empty or unstable clusters and the total number of iterations corresponds to each cluster is very less compared with the previous method.

Table 1: Number of Iterations Required for the Existing *K*-Medoid Technique

No. of clusters	***Total iterations***				
	Run1	***Run2***	***Run3***	***Run4***	***Run5***
3	138	55	49	71	103
4	89	50	120	78	98
5	50	45	45	43	69
6	90	54	56	34	30
7	94	74	75	95	30

From the above table 1 we can see that the total number of iterations corresponds to each cluster is very high when we use the existing *K*-medoid method. Also sometime it generates unstable clusters. Table 2 represents the number of iterations required for modified *K*-medoid clustering based on *KL* divergence similarity.

Table 2: Number of Iterations Required for the modified *K*-Medoid Technique

No. of clusters	***Total iterations***				
	Run1	***Run2***	***Run3***	***Run4***	***Run5***
3	13	5	4	10	20
4	7	9	12	9	7
5	15	13	32	8	9
6	20	20	25	13	17
7	9	8	9	12	20

From the table 2, it can be observed that the proposed *K*-medoid clustering results in lesser number of iteration when compared to existing *K*-medoid techniques. And we can say that the modified *K*-medoid clustering does not generate any unstable clusters.

5. CONCLUSION

Several works on clustering uncertain data are studied in detail. Clustering uncertain data based on their probability distribution similarity is more efficient. Random selection of initial medoid is the main drawback of probability distribution based clustering method But the proposed method overcome that problem effectively. The experimental results conclude that the proposed method produced good results. Researches can address the following issues to improve the performance.

- *K*, the number of clusters should be fixed beforehand.
- We can integrate *KL*-divergence also into density based method

References

[1] Dr. T. Velmurugan "Efficiency of K-Means and K-Medoids Algorithms for Clustering Arbitrary Data Points ". IJCTA, 2012

[2] Samir Anjani and Prof. Mangesh Wangjari. "Clustering of uncertain data object using improved K-Means algorithm" IJARCSSE, 2013

[3] Ben Kao Sau Dan Lee Foris K. F. Lee David W. Cheung and Wai-Shing Ho." Clustering Uncertain Data using Voronoi Diagrams and R-Tree Index" IEEE, 2010

[4] Hans-Peter Kriegel and Martin Pfeifle.” Hierarchical Density-Based Clustering of Uncertain Data” IEEE, 2005
[5] Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu.” A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise” International Conference on Knowledge Discovery and Data Mining.
[6] Bin Jiang, Jian Pei, Yufei Tao and Xuemin Lin. “Clustering Uncertain Data Based on Probability Distribution Similarity”IEEE, 2013
[7] Fernando Perez Cruz.”Kullback-Leibler Divergence Estimation of Continuous Distributions”
[8] Hae-Sang Park and Chi-Hyuck Jun.” A simple and fast algorithm for K-medoids clustering”Elsevier, 2008
[9] Mrs. S. Sujatha and Mrs. A. Shanthi Sona. ” New Fast K-Means Clustering Algorithm using Modified Centroid Selection Method” IJERT, 2013
[10] Nick Larusso.” A Survey of Uncertain Data Algorithms and Applications”. IEEE Transaction On Knowledge And Data Engineering, 2009